Video Infilling with Rich Motion Prior

Xinyu Hou[1], Liming Jiang[1], Rui Shao[2], Chen Change Loy[1]

[1]S-Lab, Nanyang Technological University; [2]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)

Introduction

Video infilling: generate visually smooth and plausible intermediate frames in between given context frames with large temporal gap.

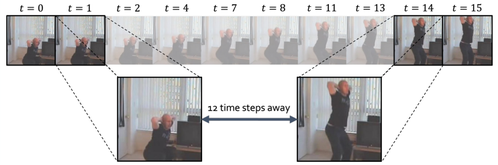

Figure 1: Illustration of video infilling context frames.

Major challenges:

- **Significant motion change** between given contexts
- **Uniform assumption** may not hold

Model should learn **robust motion dynamics**

↓

Masked Modeling!

Methodology

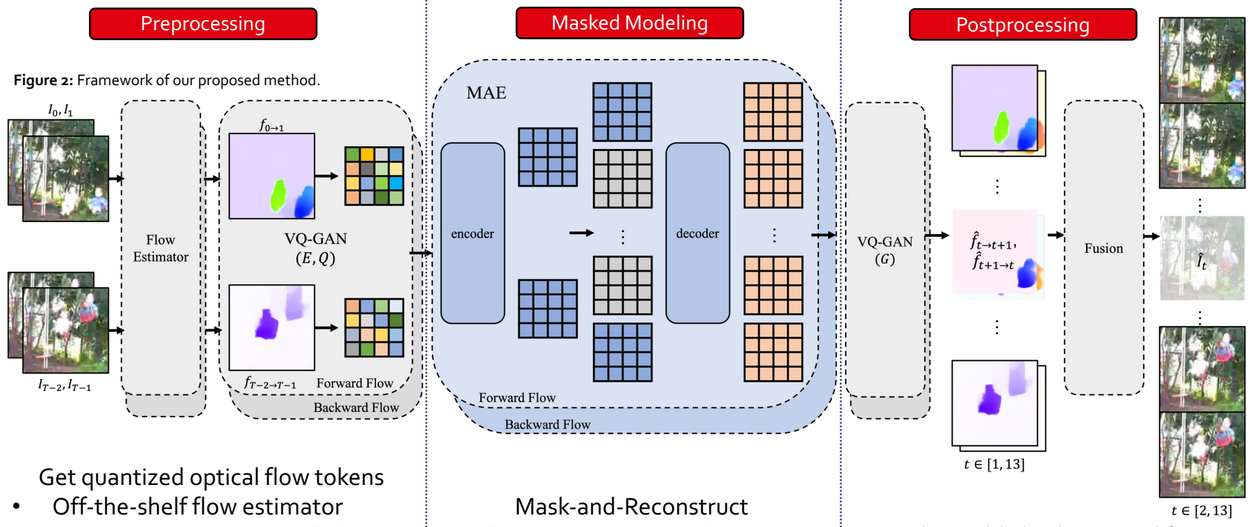

Figure 2: Framework of our proposed method.

Get quantized optical flow tokens

- Off-the-shelf flow estimator
- Pretrained optical flow VQ-GAN

Mask-and-Reconstruct

- Specially designed infilling mask

Non-learnable bi-directional fusion

Comparison with Baselines

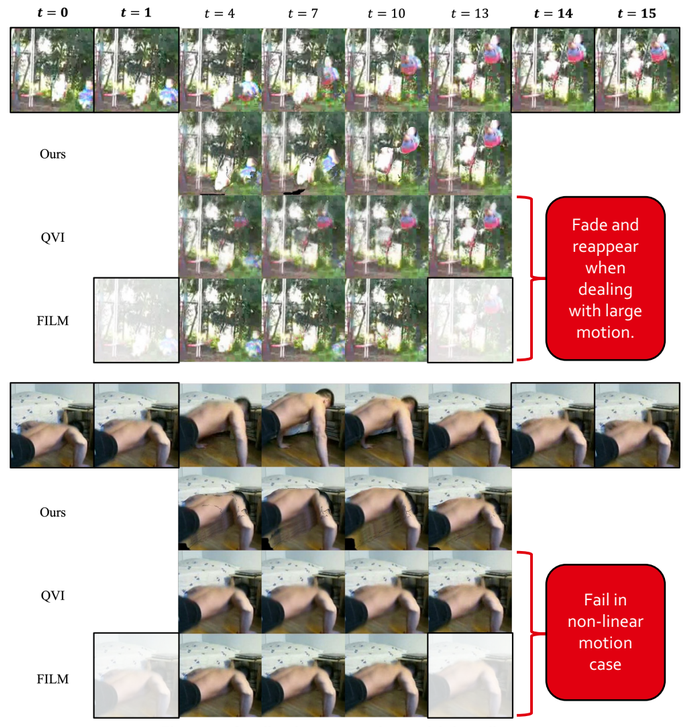

Figure 5. Qualitative comparison with baselines.

Method	By Frame			By Video
	SSIM↑	PSNR↑	LPIPS↓	FVD↓
FILM [1]	0.7571	22.85	**0.1244**	-
QVI [2]	0.7254	22.13	0.1281	1206.89
Ours	**0.8051**	**22.97**	0.1310	**1104.69**

Table 1. Quantitative comparison with baselines on the UCF101 test set..

Qualitative Results

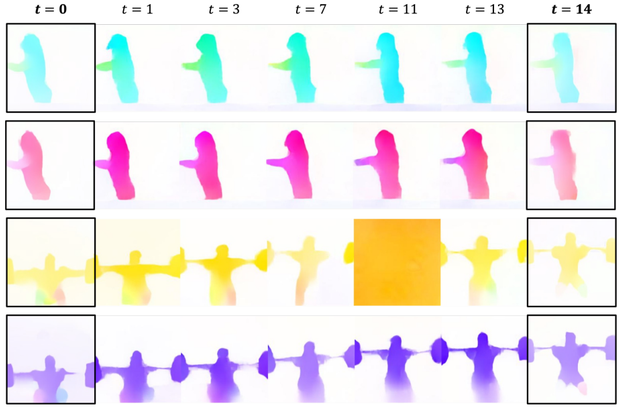

Figure 3. Optical flow prediction results. Context flows are marked with black outlines.

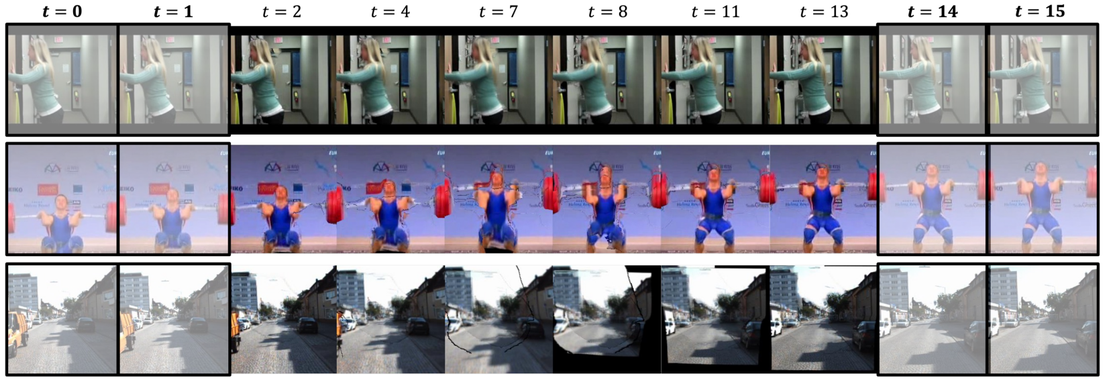

Figure 4. Pixel-level infilling results. Context frames are marked with black outlines. *Top & Middle*: same dataset; *Bottom*: cross dataset.

References:

[1] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. FILM: Frame interpolation for large motion. In European Conference on Computer Vision (ECCV), 2022.

[2] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In Advances in Neural Information Processing Systems (NeurIPS), 2019.